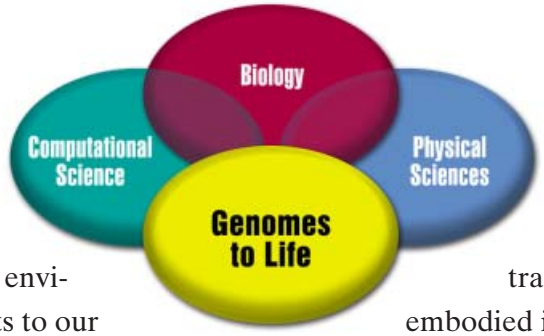
Realizing the Potential of Genome Projects:

DOE Resources and Technology Centers for Biological Discovery in the 21st Century

GENOMES to LIFE
BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES
U.S. DEPARTMENT OF ENERGY
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

Office of Science | Revised May 2003 | DOEGenomesToLife.org



The revolution in biology triggered by the human and other genome projects now promises far-reaching energy, environmental, and health benefits to our nation. Today, scientists have in hand the complete DNA sequences of genomes for many organisms—from microbes to plants to humans. For the first time, we can begin to explore the "operating systems" of life written into these genetic codes and put them to use.

To realize the full potential of genome projects we must now move beyond studying single biological "parts" (one gene or protein at a time, for example) and begin to explore how they work together to form dynamic living systems. At the leading edge of this great scientific frontier is the Genomes to Life program, whose overarching goal is to use these tools to target critical DOE mission challenges. Potential benefits of this program are enormous, with perhaps the most important being (1) clean, biology-based energy to enhance energy security; (2) reduced carbon dioxide in the atmosphere to help stabilize global climate; and (3) dramatic savings and more effective environmental cleanup.

Research centers housing a new generation of sophisticated, high-throughput technologies are required to translate the new biology, embodied in the goals of the Genomes to Life program, into a reality. Implementation of the plan outlined in this document will result in a robust and comprehensive resource base for all DOE programs, and will be a critical enabler of the national effort in 21st-century biology. This plan stems from a report approved by the Biological and Environmental Research Advisory Committee at its meeting on April 25–26, 2002.

A Systems-Level View

The Genomes to Life program reflects the fundamental change now occurring in the way biologists think about biology. The program exploits the data and extends the new paradigm of comprehensive, whole-genome biology to the next level—a whole systems approach. This requires merging concepts and technologies from the biological, physical, and computational sciences to enable simultaneous explorations of complex interactions at many levels of biological function.

Building on a History of DOE Investments

Biological Solutions for Energy Challenges

DOE investment in high-risk, high-payoff research has built a critical foundation for many national successes in the biological and medical sciences. Today, this tradition continues for the many new, high-profile programs being launched. Recent outstanding examples include the Human Genome Project, medical-imaging technologies such as the forerunners of CT and PET scans, medical isotopes for diagnostic imaging, chemical-biological sensors, and enabling the Protein Structure Initiative at NIH.

With its Genomes to Life program, DOE now leads the way in pioneering new technolgies, resource centers, and research for understanding how the individual components of life—genes and proteins—work together in complex biological systems. DOE capabilities in high-performance computing are an essential part of a new biology toolkit designed to facilitate the understanding of entire living organisms. Through multidisciplinary collaborations at national laboratories, universities, industry, and across federal agencies, DOE is poised to generate the knowledge, tools, resources, and unique facilities that become the core of research programs in the nation and worldwide.

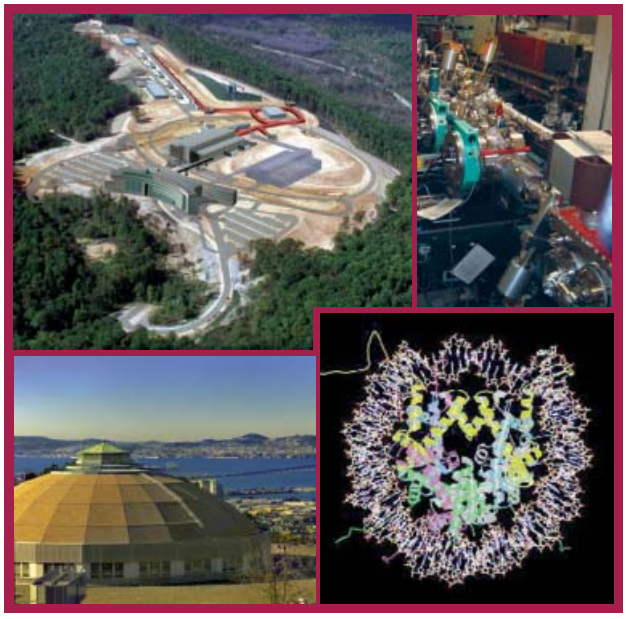
Production Sequencing Facility at the DOE Joint Genome Institute

The Human Genome Project (HGP), launched by DOE in 1987, has contributed new insights into the complexities of the genetic instruction manual—the DNA sequence—showing that large projects in biology, as in physics, could lead to new frontiers of discovery.

Because the genome contains all the information necessary for the cell to create and sustain complex living systems, genome data provides a starting point for understanding life's processes in a fundamental, comprehensive, and systematic way. DOE investments in DNA sequencing and sequencing technology helped generate this essential new knowledge base.

Spallation Neutron Source being built at Oak Ridge National Laboratory

DOE investments in synchrotron and neutron sources, although focused initially on the physical sciences, are giving new insights into protein structure and function. Today, DOE is at the forefront in developing large-scale approaches to studying all of an organism's proteins—the rapidly emerging field of proteomics. Also, DOE-supported imaging technologies, the first to see inside patients, now allow scientists to look inside individual cells to understand how proteins function together in a living organism.



A beamline at the National Synchrotron Light Source at Brookhaven National Laboratory

Molecular structure of the nucleosome core complex

Advanced Light Source at Lawrence Berkeley National Laboratory

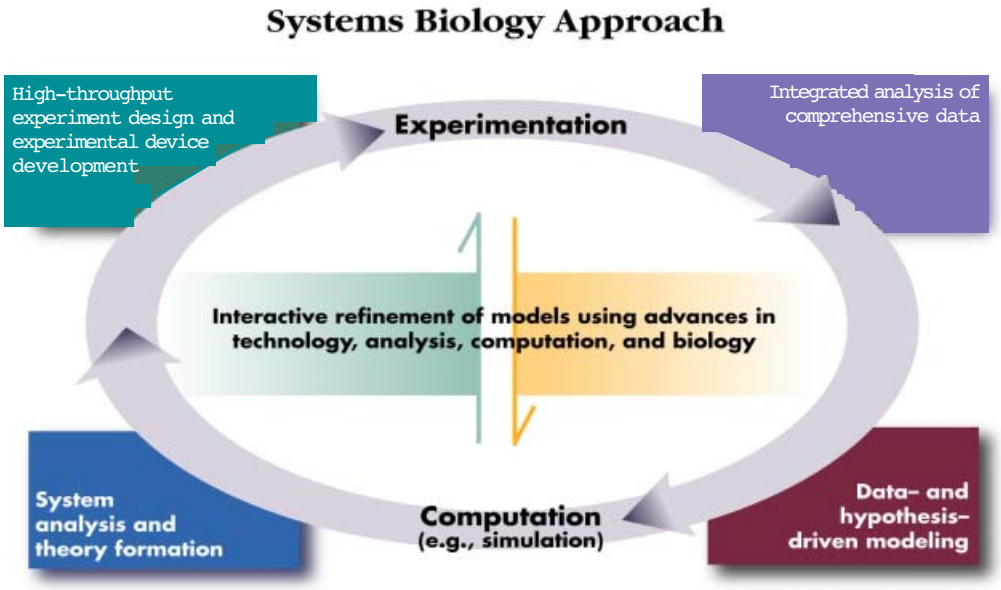
The strategy in Genomes to Life exploits the exquisite functional diversity of microbes—nature's simplest and most abundant organisms. Microbes have evolved for some 4 billion years to establish niches in virtually every environment, and have met challenges that include extremes of temperature, pressure, salinity, and high radiation levels. An understanding of genes, proteins, and their regulatory systems will enable scientists to predict how these cells function and respond to different environments. This understanding will allow us to design ways to put the biological capabilities of various organisms to work.

Technology and Computing Needs

Current state-of-the-art instrumentation and computation enable and encourage the establishment of this ambitious and far-reaching program. However, concurrent technology development is needed to reach Genomes to Life goals within the next decade. Substantial efforts must be devoted, for example, to improving technologies for characterizing proteins and protein complexes, localizing them in cells and tissues, carrying out high-throughput functional assays of complete cellular protein inventories, and sequencing and analyzing microbial DNA taken from natural environments.

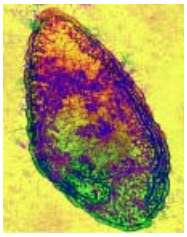
Further, the wealth of data to be collected in studies of dynamic living systems will have meaning only if it can be assimilated, understood, and modeled on the scale and complexity of real living systems and processes, a task requiring advanced computational methods and capabilities. To meet these grand challenges of the new biology, the offices of Biological and Environmental Research (BER) and Advanced Scientific Computing Research (OASCR) within the DOE Office of Science have formed a strategic alliance to lay the foundation for a large computational and mathematical infrastructure. These advances will complement experimental and theoretical biology to build a larger, intellectually richer and more agile biology enterprise. These synergies are critical for making biological knowledge widely available and stimulating new biological discovery and understanding.

GTL Program Goals



A new "systems biology" approach explores complex interactions of many levels of biological information simultaneously. Coupling experimental design and data analysis with computation will ultimately be key to the success of Genomes to Life and enable an integrated and predictive view of how cells behave and respond to environmental changes. The creation of advanced high-throughput technologies and new centers for biological computing and information management will be essential for achieving these goals.

Applications of this knowledge will provide revolutionary new approaches for harnessing the capabilities built into living systems (such as the bacterium *Nitrosomonas europeae*, at right) to meet DOE mission challenges in energy security, global climate change, toxic waste cleanup, and health protection.



Call for New Resources

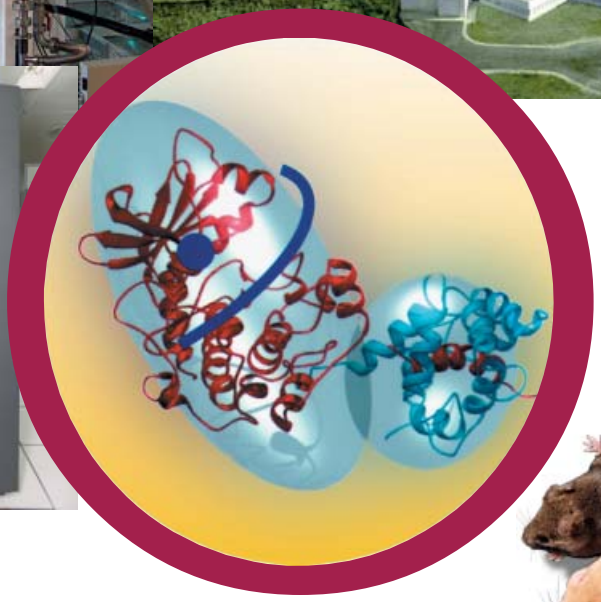
A key part of this new enterprise is to make the full suite of technologies available to individual researchers, as well as to provide access to vast new information resources that will rapidly evolve. Achieving the promise of GTL will require a philosophy that brings the life, physical, and computing disciplines together with a coherent set of goals. This enterprise must harness the unique powers and resources of the national laboratories, academia, and industry in new ways if promise is to become reality.

As the genome projects have shown, a comprehensive and high-throughput approach to biology not only provides an entirely new model and approach to understanding life but also requires innovative institutional models that go far beyond the historical single-investigator model. Critical to achieving GTL goals is genome-scale collection, analysis, dissemination, and modeling of data. Just as with the Human Genome Project and the community's production of DNA sequence, a key to GTL's success will be the generation of genome-scale data and capabilities for data management and analysis to interpret the biological "outputs" of a genome. This will enable an understanding of biological function that occurs through genetic regulation, the activities of individual molecular machines, higher-order structure and function of cells, organisms, and microbial communities. Creating new capabilities, making them widely and readily available, and using them effectively can best be done by establishing new technology and resource centers to serve the community of national laboratory, academic, and

Environmental Molecular Sciences Laboratory's 800-MHz nuclear magnetic resonance spectrometer at Pacific Northwest National Laboratory

Crystallography resources such as those at the Advanced Photon Source at Argonne National Laboratory

National laboratory supercomputing capabilities



Mouse mutagenesis capabilities

A Prototype for Success

Intricate details on the binding and dynamics of a muscle protein and its activator (circular inset) were painstakingly revealed over the course of several years, using a combination of mouse mutagenesis, crystallography, NMR, neutron scattering, and computational technologies by scientists at Los Alamos National Laboratory. Examples of these capabilities are pictured above.

New GTL technology and resource centers are necessary to provide the tools for implementing the high-throughput, comparative whole-systems approach that is a hallmark of the new biology. Broad access to a full suite of advanced and diverse technologies within an integrated computing and information environment will enable more effective research and stimulate new avenues of inquiry not possible today.

Success in these efforts will yield profound insights into how genes and proteins function together—as protein complexes (cellular machines), within cellular networks, and within a diverse community of life forms, including those with capabilities that can be harnessed for DOE applications.

industrial research users. A concerted strategy must speak to a wide range of institutional needs and requirements.

Rationale and Benefits for Technology and Resource Centers

DOE's GTL technology and resource centers will serve a wide variety of specific purposes, including the following:

- Assemble and facilitate new capabilities for high-throughput systems biology.
- Enable systems biology research to take full advantage of existing national facilities.
- Create next-generation centers for biological computing and information management.
- Advance technology development that enables the implementation of GTL.
- Facilitate the application of GTL science and technologies to specific areas within DOE's portfolio, especially specific DOE mission areas.
- Provide resources and user facilities for scientists throughout the national laboratories, academia, other federal agencies, and industry.

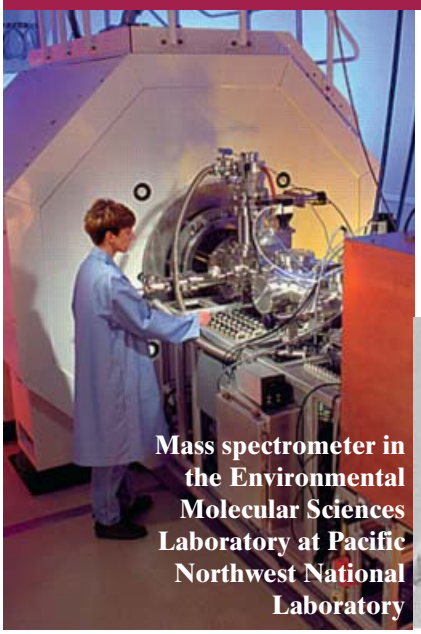
Such technology and resource centers can have many benefits.

1. They will foster new science. Providing scientists with the ability to open new avenues of inquiry will fundamentally change the course of biology in coming decades, enabling totally new kinds of questions to be asked and answered. This will attract the very best talent to the enterprise.
2. They will stimulate multidisciplinary technology development. Many technologies can be developed only in an environment of deep and broad technical and engineering resources.
3. They could be the only way to meet technology challenges of great significance to biological discovery. The synchrotron sources represent a striking example of this phenomenon for determining protein structures.
4. They can achieve economy and efficiency of scale, as illustrated in genome sequencing centers such as DOE's Joint Genome Institute.
5. They can bring together advanced and diverse technologies and programs in an integrated computing and information environment, thus providing a comprehensive infrastructure.
6. They can be the point of focus of whole new communities of scientists at the interfaces of disciplines.
7. They can foster cooperation among institutions, providing a new venue for science that transcends national laboratory, academic, and industrial boundaries.

The development of major facilities, a great success story for DOE, has led to many revolutionary avenues of research and discovery for the nation. New GTL resources will include distributed networks of resources, enhancement of existing user facilities, and new stand-alone centers for new purposes.

Existing Facilities Relevant to GTL

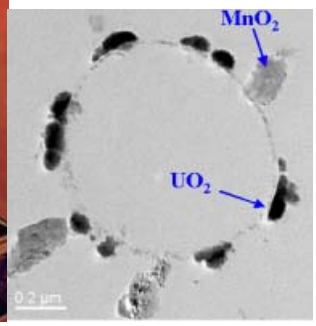
Existing DOE-supported facilities provide an important foundation for GTL to build upon. As the agency primarily responsible for creating and operating major scientific user facilities, DOE has a robust assortment of such resources that are used by a very large multidisciplinary community. A key GTL priority is to establish support, production, and other resources to ensure optimal access to and utility of GTL-related research and applications.



Mass spectrometer in the Environmental Molecular Sciences Laboratory at Pacific Northwest National Laboratory

Bioremediation of Hazardous Waste

Biotechnology offers one of the most promising approaches for remediating toxic wastes to environmentally safe levels. Exploiting the biological capabilities of microbes (e.g., *Shewanella*, right photograph) that can contain, eliminate, or decrease these materials requires a much more complete understanding of fundamental biological processes.



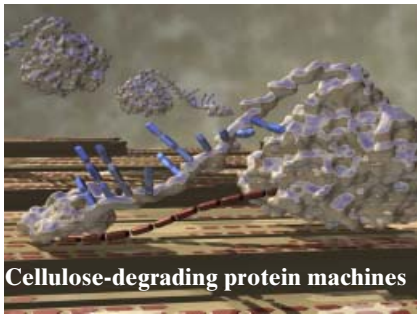
Integrating whole-proteome technologies using advanced, high-throughput mass spectrometers (left photograph) with new computational models can help generate a predictive understanding of cell behavior in dynamic settings. This knowledge can be applied to develop novel ways to enhance bacterial metabolism for treatment of hazardous waste.

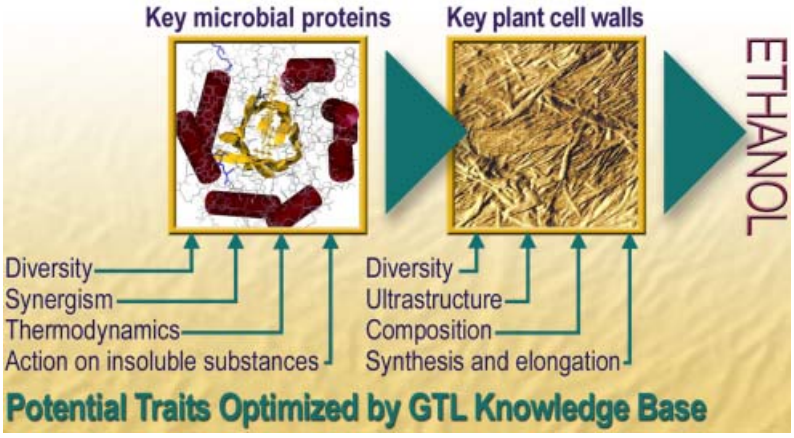
In the case of the four synchrotrons (Advanced Light Source, Advanced Photon Source, National Synchrotron Light Source, and Stanford Synchrotron Radiation Laboratory), partnership with the National Institutes of Health (NIH) and other nonfederal partners has created a very effective and large set of instruments that serve the nation's needs for doing pioneer work in structural biology. Strong programs in structural genomics are being developed at all four synchrotrons, primarily with NIH funding. These facilities form the nucleus of strong, innovative scientific programs, a phenomenon that could be duplicated in other new areas. The most pragmatic approach would be to enhance and make available existing capabilities to support GTL's structural biology aspects.

In addition to synchrotrons, existing DOE capabilities for biology include the following:

- Significant sequencing capability at the Joint Genome Institute; annotation and sequence finishing at Oak Ridge National Laboratory (ORNL), Los Alamos National Laboratory, and Stanford Medical School; and the comprehensive microbial databases at The Institute for Genomic Research and ORNL
- Leading NMR imaging and supporting isotopic labeling capabilities (PNNL and others)
- Leading mass spectroscopy (MS) capabilities for proteomic and other applications (ORNL, PNNL, and other laboratories)
- Mouse house and mammalian genetics and genomics at ORNL
- Ribosomal Database Project at Michigan State
- National Centers for High Performance Computing
- Electron microscopes
- Laser-based imaging facilities
- Neutron facilities at the High Flux Isotope Reactor Facility, Los Alamos Neutron Science Center, and the future Spallation Neutron Source

Clean, sustainable energy





Enhanced plant qualities and microbial bioprocesses can be used to generate clean, sustainable energy. Microbial protein "machines" can break down the cellulose in plant cell walls for fermentation to ethanol. Today, the process is too inefficient for commercial production. Fundamental knowledge of gene regulation and protein machines gained in GTL can be applied to develop highly efficient methods to support large-scale ethanol production and displace a significant amount of fossil fuel use.

- Technology development for genomics and proteomics, sequencing, gene-expression measurement and comparison, in several locations

In several instances, the impact of existing or expanded facilities could be significantly increased by providing for more efficient access by the wider scientific community. Where appropriate, such improvements in access should become a high-priority goal for GTL.

New Technology and Resource Centers

The creation of this new generation of technology centers offers an opportunity to leverage the best in the life sciences' individual-investigator tradition with the physical and computing sciences' tradition of creating sophisticated and technologically advanced facilities and computing infrastructure. The challenge of increasing the scale of data acquisition and experimentation and of developing technologies for working at the GTL scale requires the establishment of several types of research centers. Some of these could be fielded immediately in pilot facilities to test them against real biological problems, train a community of scientists in their use, and understand the economies of scale. This would help define the scope as well as the technical and support requirements of these kinds of facilities.

The advantages of centralized capabilities must be determined definitively. In many cases, centralization is clearly needed from the outset, but a distributed approach may have advantages for other types of requirements. The GTL program will need centers of several different types that generate experimental data and materials and promote computational data analysis. The following list illustrates areas for development and is not comprehensive.

GTL Experimental Data Centers

These are the primary centers for generating laboratory data of the types specified in the Genomes to Life goals 1–3.

- Goal 1 centers for the analysis of multiprotein molecular machines that develop and employ technologies for identifying and understanding interactions between cellular proteins and their functions. Required technologies include large-scale and high-throughput protein expression and production, sample separation and preparation, instrumentation, mass spectrometry analyses, and integrated bioinformatics and computation.
- Goal 2 centers for mapping and modeling gene regulatory networks that initially pilot and later scale up regulatory network discovery and mapping. This involves the development and integration of whole-cell proteomics capabilities, large-scale gene and protein chip analyses, comparative genomics including multispecies large-insert libraries, new methods for analyzing cis-regulatory elements in the genome, gene regulatory network bioassays, and an integrated computational and bioinformatics program.
- Goal 3 centers for the analysis of microbial growth and interaction that include chemostats and fermentor farms to study the growth and dynamics of microbial systems in pure and mixed cultures under a variety of conditions. As studies progress, these centers will develop laboratory-scale pilots for investigating various scenarios in energy production, biomass conversion, and carbon sequestration. Technologies will include microbial imaging capabilities such as atomic force microscopy; scanning electron microscopy to image live-hydrated microbes, and related high-resolution imaging technologies; microchemistry capabilities such as ion microprobe-type analyses; focused ion beam; secondary ion mass spectroscopy for high-resolution chemical mapping of intra- and extracellular enzyme complexes and cell-wall components of microbes; electron microscopy linked to electron energy loss; spectroscopy to facilitate microchemical analyses; and integrated modeling and simulation capabilities.

GTL Resource Centers

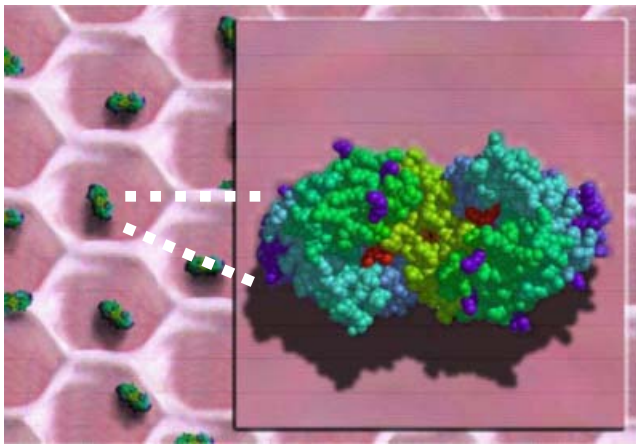
In addition to those for generating primary data related to GTL goals, additional centers are needed for foundational data and materials in support of GTL experimental data centers. These resource centers can be viewed as "pilot plants" that develop and apply important enabling technologies and key materials for GTL.

- **Production protein-isolation centers** for large-scale production of proteins essential in understanding protein complexes and protein-based materials, including composites. These centers will produce milligram quantities of thousands of proteins for use in function studies, assays, and structural analyses.
- **Centers for high-throughput proteomics** essential for a systems approach to biology. Production centers are needed to gather large-scale data about protein expression, metabolic pathways and parameters, and function. These centers will perform high-throughput, global, ultrasensitive, and quantitative measurements of RNA and protein expression and metabolic measurements; and provide informatics and computational tools to manage, analyze, and allow access to information produced.
- **Automated crystallization centers** to carry GTL-relevant proteins on to crystallization and structure determination. This is essential for proteins being studied by macromolecular synchrotron X-ray diffraction and neutron scattering. Such centers need not be located directly at the synchrotron because frozen samples are transported readily.
- **Combinatorial chemistry centers** for small-molecule–based functional genomics that

integrate the design and synthesis of novel small-molecule libraries, development of novel high-throughput biological assays, and creation of inventive strategies for identifying gene function. The rapid development of specific inhibitors of protein function, for example, can be powerful reagents in the dissection of protein function.

- **Molecular imaging centers** to develop new labeling chemistries and imaging capabilities leading to high-throughput capabilities in areas that include cryoelectron microscopy, soft X-ray microscopy, small-angle X-ray and neutron-scattering capabilities, and single-molecule detection methods that allow simultaneous imaging of several molecules and can be used to characterize the functional dynamics of proteins, including their subcellular location.
- **Intermediate-scale imaging centers** to look at many different functional states of complexes with Cryo-EM and use crystallography to reveal atomic-resolution structure. Recent ribosome studies show the value of both approaches being pursued in parallel. In the not-too-distant future, similar experiments might be possible with X rays from "fourth-generation" synchrotron light sources. Such a source (LCLS) being planned for construction by DOE could be operational in about 5 or 6 years.
- **Mouse centers** for studying the functions of biological systems in vivo in mice. The advent of genomic sequence information and revolutionary genome-based techniques has given a new importance to such studies. The generation of mice is primarily relevant to Goals 1 and 2 of GTL. Center needs include advanced technologies for manipulating mouse genes, such as the new Lenti-virus vectors that appear to simplify the process of making transgenic animals; and the ability to temporally manipulate specific gene expression and to image specific gene products in real time in live animals. Additionally, user facilities should be supported for the systematic production of transgenic and gene-replacement animals and to facilitate mutagenesis. For example, the new ethylnitrosourea (ENU) mutagenesis projects can mutagenize any and all genes in vivo, examine the phenotypic effects by high-throughput screening, and rapidly map the causative genes. The last can be done only in large central facilities.
- **Centers for the analysis of nanoscale biological structures** to determine the connection between genes and molecular machines and materials, a major opportunity for the BER program. The burgeoning field of nanoscience and nanotechnology offers a multifaceted opportunity when coupled with the revolution in biology. All of life's processes are based on

Biology + Nanotechnology for Advanced Applications



Engineered protein machines can be embedded in synthetic nanomembranes. These may one day be used to break down cellulose for more efficient ethanol production, produce hydrogen for fuel cells, or remediate waste streams.

nanomachines and phenomenology at the nanoscale and are encoded in the genomes of all organisms. A pilot center with emphasis on the genomic analysis of biological materials would be very timely.

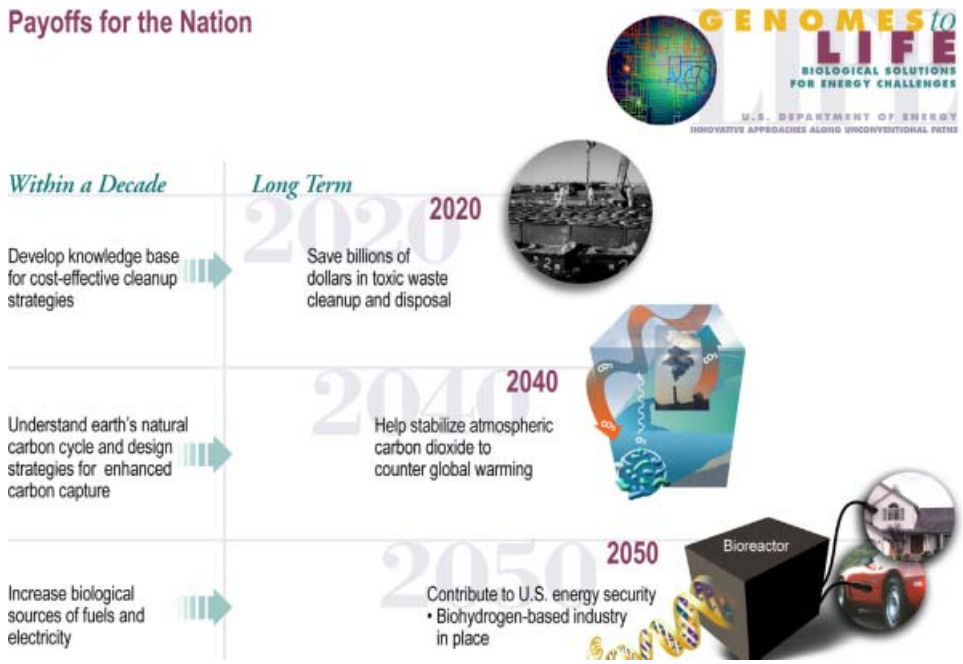
- **Large-scale DNA sequencing centers** for selected elements of whole genomes. This will be an essential ingredient in regulatory-network discovery and other GTL goals for years to come. Centers specifically designed to provide diverse researchers with access to high-efficiency, large-scale sequencing is crucial to maximizing GTL discovery and DOE's sequencing capacity.

Computational Capabilities

Linking the national laboratories' computational capabilities to the day-to-day operation of Genomes to Life research centers is one key to making these facilities truly powerful and unique. The computational biology roadmap, being developed jointly by BER and OASCR, projects a computational biology enterprise that includes the development and use of new computing and information technologies for handling, storage, and retrieval of data and knowledge; an aggressive program of mathematics and computer-science research focused on methods for modeling complex biological systems and comparing models to experimental data; and an appropriate computational infrastructure for these activities. The amalgamation of these efforts requires additional funding and mathematical analysis, software, and algorithm development dedicated to these biological problems.

Computational resources are needed in three specific roles:

- **Compute resources** to provide, maintain, and support the GTL compute and networking hardware and software infrastructure. These resources would support high-performance computers and networks tuned to the needs of biology that would allow users to effectively access and run applications on this distributed infrastructure.



- **Data resources** to bring together related data from various GTL centers and organize the information in a fashion useful to the research community. GTL experimental data centers will generate huge amounts of data (petabytes in the near future) that must be managed both locally and centrally.
- **BioTools library resources** to organize, maintain, and support software and other applications for use on a variety of hardware environments. GTL will require many new types of computational analysis, modeling, and simulation for biology and, by its nature, will need an immense software-development effort. Resource centers must support central tool libraries, tool portability, and software sharing across the GTL enterprise.

Conclusions

DOE resource and technology centers for GTL will provide the capability to bring together the biological, physical, and computing sciences to enable a fundamental understanding of life. BER played an instrumental role in bringing about the genome revolution. With these new centers and scientific resources, BER and OASCR are ready to make equally seminal contributions to the new era of systems biology that will yield innovative biological solutions to DOE mission challenges.

U.S. Department of Energy Office of Science

Marvin Frazier
Office of Biological and Environmental Research (SC-72)
301/903-5468, Fax: 301/903-8521
marvin.frazier@science.doe.gov

Gary Johnson
Office of Advanced Scientific Computing Research (SC-30)
301/903-5800, Fax: 301/903-7774
gary.johnson@science.doe.gov

Office of Science: www.sc.doe.gov
Genomes to Life Program: DOEGenomesToLife.org